Grenze International Journal of Engineering and Technology, June Issue

A Comprehensive Survey on AI Hallucination Detection and Trust Evaluation in Large Language Models

Nithin S[1] and Tejaswi K[2]
[1-2]Department of Computer Application, JSS Science and Technology University Mysore, India
Email: niithins14@gmail.com, tejaswi@jssstuniv.in

***Abstract*— Today, applications are powered by large language models (LLMs) like Claude, GPT-4, LLaMA and Gemini in healthcare, education, law, finance and science. They also continue to hallucinate, making things up, giving out 'garbage data' or just being completely wrong. This is a reduction in user trust and difficult for safe deployment. This survey summarizes the research on hallucination detection and trust evaluation in the period of 2020 to 2026. We propose a five-class hierarchical taxonomy of hallucination, classify over eighty detection methods into four paradigms (internal, external, hybrid, and human-in-the-loop) and establish a trust evaluation framework with eight dimensions. In addition, over 15 benchmarks are explored, benchmarks metrics are compared, and nine open challenges and ten directions for future work are presented. Lastly, we sketch the Hallucination-Aware Trust Architecture (HATA), a modular pipeline consisting of detection, verification, scoring and explainability for more trustworthy deployment.**

***Keywords*— Hallucination detection · Large language models · Trust evaluation · Fact verification · Retrieval-augmented generation · Benchmarks · Explainable AI · Generative AI reliability.**

I. INTRODUCTION

In the last decade, AI has moved from narrow AI, which focuses on specific tasks to generative AI, which is a generative foundation model that can create human-like texts, The materials are made up of code and multimedia content. Transformer-based LLMs like GPT-4, PaLM-2, LLaMA-2, Gemini, and Claude are now capable of competently answering questions, Writing code, summarizing and reasoning about scientific problems. As these As systems are plugged in to commerce and society, so are the stakes. The global LLM market is projected to grow to be as important – becoming just as significant. Increase to USD 140 billion by 2030 from USD 6.4 billion in 2024. The same mechanisms which enable these models to do well also make them their Achilles' heel: hallucination, the production of convincing and confident speech that is simply wrong. The cost will depend on the incident's location. A hallucinated When drugs interact with each other, it can be risky for the patient. Any bogus legal citation can lead to As in the 2023 Mata v. Avianca case, attorneys for a doctor have been found guilty of professional penalties. Supplied case law that was non-existent. In financial matters, the fake takes on the form of money, and in science, the faked reference disturbs the integrity of the record.

Grenze ID: 01.GIJET.12.2.27_25

The challenge is not a new one, but there is no consensus of a definition of what constitutes a challenge, nor as to a common approach for measurement, and no A universal solution that will help to ease the challenge.

A. Motivation and Contributions

Existing studies mostly focus on either LLM hallucination or AI trustworthiness as standalone tasks, with few studies comparing and contrasting benchmarks, datasets, metrics, and methods together. This survey aims to fill that gap, and has five major contributions to make:

Comprehensive taxonomy. A hierarchical structure consisting of five types of hallucinations, each with their causes traced back to the data, model, prompt and retrieval levels.

- Detection analysis. More than 80 detection methods are compared and quantified, in four paradigms.
- Trust framework. A trust evaluation architecture designed in 8 dimensions, integrated in the pipeline of HATA.
- Benchmark review. Over 15 hallucination datasets including TruthfulQA, HaluEval, FEVER, and FActScore..
- Research roadmap. There are nine challenges open for the solution and ten future directions for further research, all of which are associated with gaps we find along the way.

B. Survey Methodology

The search was conducted using a protocol inspired by the PRISMA guidelines. The search strings used were "LLM hallucination detection" and "language model trust assessment" and literature was collected from IEEE Xplore, ACM Digital Library, SpringerLink, arXiv and Semantic Scholar. Works that were accepted had to be: peer-reviewed or be a heavily cited preprint (at least 50 citations); centered primarily on the theme of evaluating trust or hallucination; published between January 2018 and March 2026. 215 works were then selected for close analysis after discarding the duplicates and making quality cuts. The remainder of the paper proceeds with the fundamental and taxonomy (Sect. 2), detection methods (Sect. 3), trust frameworks (Sect. The process in 4) is the benchmarking and measuring process. 5), comparative analysis (Sect. 6), applications (Sect. Opening up research questions and future directions (Sect. We also tried the proposed HATA framework (Sect. 9), and the conclusion.

II. FUNDAMENTALS OF AI HALLUCINATION

A. Definition

First, "hallucination" was coined in image processing in the context of artifacts in super-resolution. Maynez et al. [3] provided its formal definition, and used it for summarization systems that generate statements that were not contained in the source. Since then many other definitions have been coined. We have taken a broadly drafted definition for this survey. An AI hallucination is defined as any part of the model's output that is: (a) incorrect in relation to real world facts; (b) inconsistent with the model's previous statements; (c) not supported by or contradicted by the context the model was provided with; (d) outright fabricated, e.g., a fake citation; or (e) cross-modally inconsistent in multi-modal systems..

B. Root Causes

Hallucinations are of four causes. The data problems are noisy web-scraped corpora, skewed topic distributions, fixed cutoff of knowledge, and thin coverage of rare, long-tail entities. At the model level, early errors compound via autoregressive decoding, the probability of a token is distributed over the knowledge, memorization and generalization are opposing forces, and RLHF can reward answers that are only confidently correct, but not really. Prompt-level: Ambiguous queries, adversarial queries, Prompts exceeding context window. A fourth set comes to the surface when the retrieved documents are on-topic and factually incorrect, or when the retrieved evidence is contradictory, or when retrieval fails, and the model turns to "stale" parametric knowledge.

C. Taxonomy of Hallucination Types

There are five main categories of hallucinatory experiences with subcategories for each. The entire structure that will be used in the discussion below is outlined in Figure 1.

- Factual Hallucination. Statements that contradict what is verifiable world knowledge. The subtypes activate the wrong thing: entity hallucination refers to the misidentification of persons, organisations or places; temporal hallucination refers to errors in dates or sequences; numerical hallucination pertains to errors in

quantity or statistics; causal hallucination refers to the connection of wrong cause to wrong effect. This is the most researched class.

- Logical Hallucination. Errors in reasoning, in which the conclusion is not implied by the reasons; faulty deductions, self-contradiction within a single response, invalid analogies. They can be particularly harmful in multi-step reasoning and mathematical situations where the first error cascades through the subsequent steps.
- Fabricated Reference Hallucination. Academic citations, case law, news articles or product reviews that the model creates and which don't exist. Mata v Avianca was this year's front page story in this category. Citation-hallucination rates reported range from 30% to 90%, and are even poorer in the legal and medical domains.
- Contextual Hallucination. In grounded tasks, the model is given a source document to work with, and errors are recorded in these tasks. The intrinsic version does the opposite, containing information which is not in the source, but may be true by itself. These are well-documented in summarizing and grounded questions.
- Multimodal Hallucination. Failure in visual, sound and video-language models. Object hallucination reports something that isn't in the picture; attribute hallucination reports the wrong color, size, position; relational hallucination reports a relationship that is incorrect in space or meaning; and cross-modal inconsistency reports text that contradicts the image or sound or table input it should correspond.

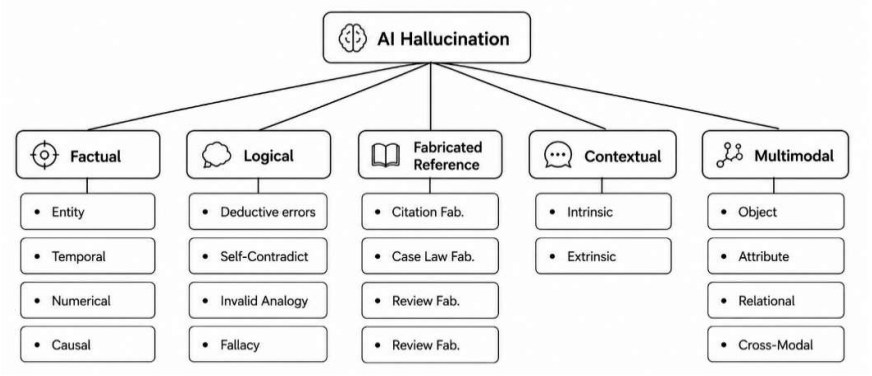

Fig. 1. Hierarchical taxonomy of AI hallucination types and their subtypes

III. Taxonomy of Hallucination Detection Methods

There are four major paradigms of detection approaches, as illustrated in Figure 2. Each represents a specific way of establishing the association between the detector and the model it is monitoring, each has different tradeoffs for precision, latency, and out-of-band resources.

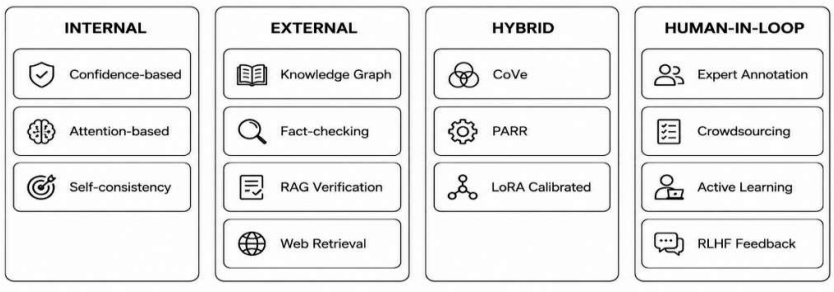

Fig. 2. Taxonomy of hallucination detection methods organized by paradigm

A. Internal Detection Methods

Internal methods are used to read signals within the model itself. They do not require external information and are suitable for real-time applications. Confidence-based detection substitutes token probabilities and Shannon entropy for factual certainty; Kadavath et al. [9] showed that calibrated probabilities are correlated with factual accuracy in moderate ways, and verbalized confidence elicitation [24] showed that it brings the model closer to

what it is actually accurate about. Attention-based detection monitors attention diffusion and cross-attention to detect generation that is only loosely related to the source tokens [21,22]. Self-consistency approaches differ: SelfCheckGPT [5] randomly selects multiple outputs, compares them to BERTScore, and identifies the inconsistent outputs without relying on any external data, and chain-of-thought consistency [11] eliminates logical errors by taking the most consistent output across reasoning chains.

B. External Verification Methods

External approaches validate claims to authoritative resources. They are willing to purchase higher precision for factual hallucinations, but they are willing to pay for it in terms of latency, and the need for a good corpus. Knowledge-graph verification enhances the accuracy of the Knowledge Graph by extracting the subject-predicate-object triplets from the text and matching them with the KGs, such as Wikidata, DBpedia, or biomedical ontologies, and Lin et al. [12] achieved 78% precision using this approach with the TruthfulQA data set. An output is segmented into atomic facts, evidence supporting these facts is retrieved, and natural-language inference is applied; FActScore [6] achieves 82.1% accuracy for the task of biographical generation by simply counting the proportion of atomic facts supported by Wikipedia. Retrieval-augmented verification scores faithfulness as a measure against the documents retrieved, and the RAGAS framework [23] combines together context precision, recall and faithfulness.

C. Hybrid and Human-in-the-Loop Methods

Hybrid approaches combine internal uncertainty signals and external verification, typically only the more uncertain generations are sent down the costly verification path. Chain-of-Verification (CoVe) [7] models make use of a chain of verification questions to get an 18% accuracy gain on TruthfulQA, while RARR [8] attaches retrieved web evidence to the final output to correct hallucinations afterward (+6.4 BERTScore F1). Despite all this, in the high-stakes environment, human judgment still reigns supreme. They rely on expert annotation (inter-annotator agreement score between Krippendorff's $\alpha = 0.6–0.8$); active-learning queues which only forward the uncertain outputs to reviewers; and RLAIF-style feedback fine-tuning which prunes the hallucination rate over successive iterations.

IV. TRUST EVALUATION FRAMEWORKS

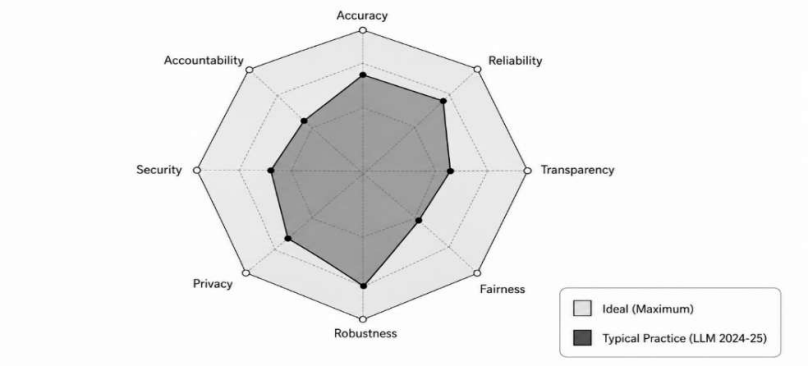

Fig. 3. Eight-dimensional trust evaluation framework. Blue polygon: typical frontier LLM; dashed outline: ideal trustworthy AI.

A. Defining Trustworthy AI

Trustworthiness is about much more than accuracy; it takes in the whole sociotechnical setting a model is deployed into. A set of multi-dimensional requirements is proposed by the EU High-Level Expert Group on AI (2019) and the NIST AI Risk Management Framework [18] — robustness, transparency, fairness, and accountability. LLMs make this more challenging, as their behavior is random and emergent. A model can have a very good benchmark and still hallucinate at alarming rates in specific areas, under challenging conditions, or in the presence of users who are unable to verify what it says.

B. Eight-Dimensional Trust Taxonomy

Figure 3 defines an 8-dimensional trust taxonomy, which we set out. The radar chart also illustrates an average frontier LLM side by side with the ideal of a fully trustworthy system, with the largest disparities appearing in the areas of explainability, transparency, and accountability. The eight dimensions are accuracy (percentage of

times that output is factually correct), reliability (stable output message when paraphrased), transparency (complete documentation), explainability (good post-hoc explanations), robustness (withstood prompt injection and OOD inputs), fairness (even
error distribution across groups), security (resisted prompt injection and OOD inputs), and accountability (audit trail and human oversight).

C. Trustworthiness Index

The various dimensions of trustworthiness are combined in a single index: TI = w1·AFA + w2·(1−ECE) + w3·Reliability + w4·Fairness + w5·Robustness, where the weights depend on the application and sum up to one. Hallucination is the most frequent loss of trust by LLM, so the hallucination rate is one of the most direct and quantifiable signals of loss of trust. It is possible to consider a whole-system assessment when the two are treated together, as is done in the HATA framework of Section 9.

V. Datasets, Benchmarks, and Evaluation Metrics

It is impossible to take a serious look at any datasets that form the benchmarks for assessment without taking a second look at them. The size of the major benchmarks is such a wide disparity that it is easy to see in Figure 4, and is summarized in Table 1, which shows for comparison the domain, hallucination type, scale, and main limitation of each benchmark.

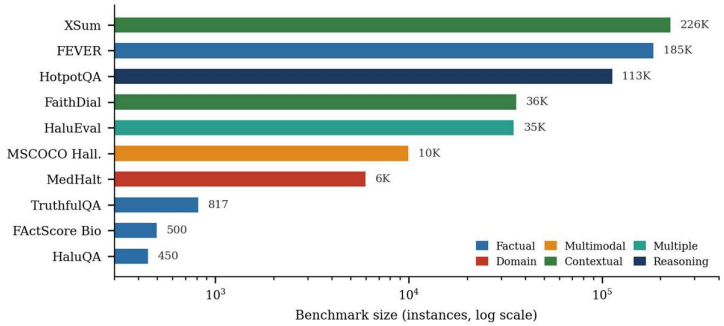

Fig. 4. Major hallucination detection benchmarks by size (log scale), colored by hallucination type.

Table I. Comparative overview of major hallucination detection benchmarks

Dataset	Domain	Hall. type	Size	Key limitation
TruthfulQA	Multi-domain	Factual	817 Q	Small scale; limited domains
HaluEval	General NLP	Multiple	35K pairs	Synthetic; may not reflect real outputs
FEVER	Fact verification	Factual	185K claims	Wikipedia-only; domain limited
HotpotQA	Multi-hop QA	Reasoning	113K Q	No explicit hallucination labels
XSum	Summarization	Contextual	226K articles	Lacks fine-grained annotation
FaithDial	Dialogue	Contextual	36K turns	Dialogue-specific only
FActScore Bio	Biography	Factual	500 persons	Person-centric only
HaluQA	Chinese LLMs	Factual	450 Q	Chinese language only
MedHalt	Medical	Domain-specific	6K Q	Requires domain expertise
MSCOCO Hall.	Visual QA	Multimodal	10K images	Image only; no text consistency
TofuEval	Meeting summ.	Contextual	1.5K summ.	Specialized domain
SummEdits	Summarization	Ctx./Factual	2.6K docs	Limited to summarization

A. Coverage Strengths and Critical Gaps

The existing benchmarks address factual hallucination in English fairly well, particularly for closed domain question answering (TruthfulQA, FEVER) and summarization (XSum, FactCC), and multi-hop question answering (HotpotQA) allows for evaluating cross-hallucination. However, there are five gaps that are hindering this field:

Multimodal coverage. Nearly no one provides any coverage of audio-text, video-text, or table-text hallucination.

- Low-resource languages. Almost all benchmarks are in English, despite the fact that success with this test is known to decrease in low-resource languages.

- Real-time knowledge. No benchmark tests hallucination in time-sensitive domain, where the ground truth is changing.
- Agentic tasks. None of them reflects how hallucinations overlap on the many steps of an agent's execution trace.
- Joint calibration. One measure of factual accuracy, and one of calibration, at the same time is impossible, and that prevents an integrated view of trust.5.2 Detection Metrics

The usual binary-classification measures are used for detection systems - precision, recall, and F1. High precision systems are preferred in user-facing systems because there will be fewer false alarms, while high recall systems are preferred in safety-critical systems because there will be fewer hallucinations. The Hallucination Rate (HR) is the proportion of hallucinated claims to all claims and ranges widely, from 2.1% (GPT-4 with RAG on medical QA) to 67% (GPT-3.5 on scientific citations). Atomic Fact Accuracy (AFA) is even more granular, breaking the response into atomic facts, and reporting the percentage accuracy of each individual fact itself, as opposed to the document-level accuracy in long-form text.

B. Trust Metrics and Human Evaluation

Essentially, ECE quantifies the accuracy of a model's confidence bands by comparing them with the frequency of successful predictions: a model with an ECE below 0.05 is well-calibrated, while frontier LLMs typically fall in the range of 0.10–0.25 unless they have been explicitly calibrated. The Reliability Score is calculated as the percentage of paraphrased response pairs that have a BERTScore F1 similarity above 0.85. Once automated metrics are exhausted, the metric is rated by humans on a 1–5 scale with the target $\alpha \geq 0.6$ as well as helpfulness, consistency and citation accuracy. Expert review continues to be the best arbiter of merit for domain-specific hallucinations in medicine and the law.

VI. Comparative Analysis of Existing Approaches

Table 2 lines up a set of representative detection methods by type, dataset, metric, performance, and main limitation. There are a couple of trends that cross over the paradigms and model generations.

Table II. Comparative Analysis of Representative Hallucination Detection Methods (2023–2024)

Method	Year	Type	Dataset	Metric	Perf.	Key limitation
SelfCheckGPT	2023	Internal	WikiBio	AUC-PR	0.81	Requires multiple samples
FActScore	2023	External	Bio gen.	Atom. acc.	82.1%	Wikipedia-only
CoVe	2023	Hybrid	TruthfulQA	Accuracy	+18%	Computationally expensive
RARR	2023	External	BEGIN/FEVER	BSF1	+6.4	High latency; web reliance
Sem. Entropy	2023	Internal	TriviaQA	AUROC	0.79	Requires decoder access
RefChecker	2024	Ext. KG+NLI	Multi-domain	P/R	88/79%	Domain-limited KG
RECALL	2024	Int.+RAG	HaluEval	F1	85.7%	RAG latency overhead
HaloScope	2024	Int. activ.	TruthfulQA	AUROC	0.85	Architecture sensitive
MedHallMark	2024	Domain-spec.	MedHalt	Accuracy	79.8%	Medical domain only
FactScore-Zero	2024	Internal	Biomed. QA	AFA	76.4%	Calibration gap, rare facts

A. Discussion

Pre-LLM fact-checkers were rule-based with clear and simple logic and were easy to operate and inexpensive to execute, but they failed to handle paraphrasing and could not access long-tail knowledge. The transformer classifiers, such as BERT, RoBERTa and DeBERTa, dominate the leader boards on the tight benchmarks like FEVER, but drop badly when they venture beyond the training domain. While retrieval augmented detection performs well when a reliable corpus is available, it does poorly on multi-hop reasoning, claims about sparse entities, and procedural/causal statements that are not easily checked in the text. A more promising recent path is the "agentic" one, where the task of verifying is broken into specialized agents to work on, like FacTool [25] and multi-agent debate [26], which is appropriate for the compound hallucinations of long-form text, and remains modular and feasible in computational terms.

VII. Application Domains and Real-World Impact

The ranges of hallucination rates as reported in the main application areas are shown in figure 5. It demonstrates the severity of the unabated issue, as well as the amount of ground that can be recovered through retrieval-augmented verification.

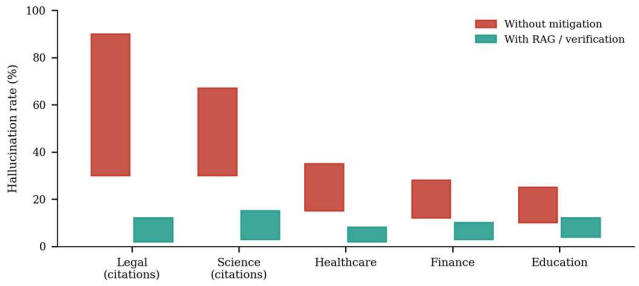

Fig. 5. Hallucination-rate ranges across application domains, with and without mitigation. Bars show documented minimum–maximum rates.

Healthcare carries the highest stakes. Unmitigated medical outputs hallucinate at rates of 15–35%, fabricating drug dosages and clinical trials that were never run, but GPT-4 with RAG brings medical-QA rates down to 2–8%, and FDA guidance on AI/ML-based Software as a Medical Device requires ongoing validation. Law has tightened up since Mata v. Avianca: citation hallucination now has to be driven close to zero by checking every citation against authoritative databases such as Westlaw and LexisNexis. In finance, domain-specialized models like BloombergGPT and FinGPT cut the error rate but still stumble on rare instruments, and in scientific research, cross-referencing against Semantic Scholar and PubMed is a basic safeguard. Education is subtler: an AI tutor that hallucinates can quietly bake misinformation into what a student learns, so detection there has to be context-sensitive and correct errors without breaking the flow of the lesson.

VIII. Open Challenges and Future Directions

A. Open Challenges

For all the progress, a handful of tangled challenges still mark the frontier. The most basic is definitional. Without an agreed, operationally precise definition of hallucination, benchmark results cannot really be compared, so a formal, hierarchical ontology of hallucination types is the groundwork everything else depends on. Benchmark gaming and contamination make this worse: once test sets show up inside internet-scale training data, reported scores inflate, which is the argument for dynamic, community-governed benchmarks that are harder to contaminate. Detection itself travels badly, too. A detector tuned on one domain reliably does worse on the next, so efficient domain-adaptive architectures stay near the top of the list, and multimodal detection across vision-, audio-, and video-language models is still barely charted, because catching a cross-modal inconsistency means reasoning over several modality-specific representations at once.

A second group of problems is about deployment. Latency bites hardest: the best external verification adds anywhere from two to thirty seconds, while production systems want a signal in under a hundred milliseconds. Long-form trust is just as unsettled, since building a coherent trust score across a long output, a multi-turn dialogue, or an agent's execution trace is much harder than scoring one claim in isolation. Three more challenges go straight to safety and fairness. Detection needs to be explainable, telling a user which span is hallucinated, why, and how to fix it, which is non-negotiable in clinical and legal settings. Robustness needs a game-theoretic footing, so a detector can hold up against prompts engineered to provoke hallucination. And equity is a real concern, because hallucination rates run noticeably higher for low-resource languages and marginalized knowledge domains, which turns into reliability gaps that track inequalities already out there.

B. Future Research Directions

The challenges suggest ten priorities for future work. The first several are architectural. Trust-aware architectures would incorporate calibration-aware training targets which would include penalties for hallucination when there is high confidence, along with machinery based on the Bayesian or conformal-prediction framework that would explicitly maintain uncertainty. With self-healing AI, frameworks, such as Reflexion [16] and CRITIC [17] would be taken to multi-step reasoning and multimodal tasks, detecting and rectifying errors in just one generation turn with formal guarantees of reliability. Explainable detection should provide a structured output with a name of the hallucinated span, a likely cause for it, and a corrected version, along with evidence. Multi-agent verification networks could be built to divide the load between the specialist agents for the entity and temporal and citation verification tasks, and there could be rules for resolving disputes between the agents. The other directions are related to infrastructure and governance. Graph-based trust models would pass trust down knowledge graphs that would dynamically change as they are updated and would resolve

contradictions when they occurred. Federated trust evaluation would combine the privacy preserving detection feedback from different deployments across the healthcare and legal sectors, without anyone sharing private information. Neuro-symbolic verification would be the linking of informal natural-language statements to formally verifiable logical statements, with the assistance of theorem provers and constraint solvers. Real-time trust scoring would exceed the 100 millisecond time budget with distilled light detectors, speculative decoding with verification folded in and with approximate nearest neighbor evidence retrieval. But under all of these lie two others: (1) longitudinal monitoring, monitoring the rates of hallucination as knowledge drifts, and models are updated, and alerting when a threshold is crossed; and (2) community-led standardization, establishing living standards of definitions, benchmark protocols and evaluation criteria, so that studies can finally be compared on equal terms, through bodies such as IEEE, NIST, and ISO.

IX. Proposed Framework: Hallucination-Aware Trust Architecture

Based on the findings of this survey, we propose a modular six-stage pipeline called Hallucination-Aware Trust Architecture (HATA) that preemptively runs over an LLM's output before it reaches the user. HATA integrates detection, verification, scoring and explainability into a single workflow, with the possibility of replacing any of these steps with improved methods without a hassle. The entire pipeline is depicted in Figure 6, with the feedback loop allowing it to continually adapt.in Figure 6, with the feedback loop allowing it to continually adapt.

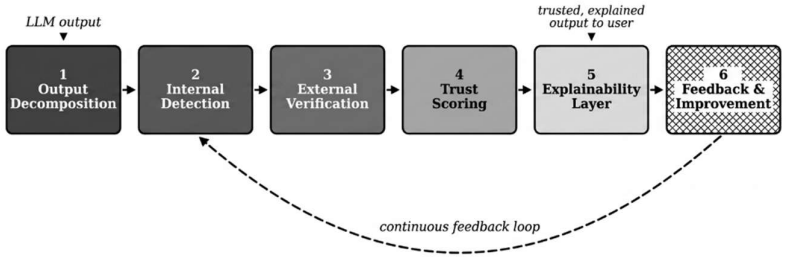

Fig. 6. The HATA six-stage pipeline. The dashed teal arc represents the continuous feedback loop enabling iterative improvement.

A. Stage Descriptions

There are 6 stages to the HATA. In Stage 1 (output decomposition), the raw output is decomposed into claim units, each of which is marked as factual, reasoning, or citation, and its contextual dependencies are noted, based on the FActScore decomposition protocol. Stage 2 (internal detection) takes approximately 50 milliseconds to score each claim based on token-confidence, entropy, and self-consistency signals calculated across 3 to 5 fast paraphrases, marking the high-confidence claims as passed for now, and passing the uncertain claims to Stage 3 (examination). Stage 3 (external verification) takes the above ambiguous claims and verifies them three times: first via a domain-specific knowledge graph, second via a web-retrieval system that rates the credibility of the source, third via a retrieval-augmented natural-language-inference (NLI) classifier that returns an entailment, contradiction, or neutral verdict. The final three steps translate the verdicts into something that can actually be trusted by the user. The verified claims are then incorporated into a document-level Trustworthiness Index (Stage 4: Trust scoring), customized to the deployment domain and displayed via a typical trust dashboard. Stage 5 (the explainability layer) processes each hallucination flag and adds a structured message: Offending span, evidence-based suggestions for correction, confidence level rephrased to the user's expertise level. Stage 6 (feedback and continuous improvement) rounds the loop back to the start, returning user corrections and expert annotations to the system to fine-tune lightweight LoRA detection adapters, update the knowledge-graph entries, and recalibrate the trust weights to adapt to shifts in the domain and new patterns of hallucinations without a full retrain.

B. Design Principles

HATA is based on 5 principles. It is modular – any stage can be replaced without having to replace the others. It is latency aware: Stages 3 through 5 are only triggered if necessary, while Stage 2 always runs, thus controlling the accuracy / speed trade-off. It is domain agnostic and trust weights and verification sources can be configured on a per-deployment basis. It is clear, adding a human-readable trust report to each output. It's also always learning, because the feedback it receives allows it to improve over time, without having to retrain.

X. Conclusion

This survey summarizes the research on the detection and evaluation of trust in AI hallucination in LLMs. Hallucination is not a single entity. It is a multi-faceted failure, encompassing inaccuracies in the facts, omissions of logic, invented references, violations of context, and the contradictions in one mode of perception with another, and every type of failure mode has its own mechanism and its own solutions. From simple perplexity heuristics to hybrid architectures, combining internal uncertainty estimates with external knowledge checks, these have progressed to AUCs of 0.80–0.90 on controlled benchmarks. However, performance in the real world is much lower than on domain-specific and adversarial queries. Despite the progress, there remain numerous open challenges for multimodal, multilingual, and agentic hallucination, and for evaluation of trust (which yet requires a many-dimensional framework that connects hallucination-specific metrics to trustworthiness, fairness, explainability, and accountability). The nine challenges and ten directions outlined here are intended as a guide for the community and the Trustworthiness Index and HATA architecture that we propose are steps toward collectively evaluating all of this. A correct hallucination detection and trust evaluation, robust, transparent and fair, is no longer an academic nicety but a societal need as LLMs enter into consequential decisions in healthcare, law, education and science.

References

[1] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Comput. Surv. 55(12), 1–38 (2023). https://doi.org/10.1145/3571730

[2] Huang, L., Yu, W., Ma, W., et al.: A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv:2311.05232 (2023). https://doi.org/10.48550/arXiv.2311.05232

[3] Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On faithfulness and factuality in abstractive summarization. In: Proc. ACL 2020, pp. 1906–1919. ACL (2020). https://doi.org/10.18653/v1/2020.acl-main.173

[4] Bang, Y., Cahyawijaya, S., Lee, N., et al.: A multitask,multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In: Proc. IJCNLP-AACL 2023, pp. 675–718. ACL (2023). https://doi.org/10.18653/v1/2023.ijcnlp-main.45

[5] Manakul, P., Liusie, A., Gales, M.J.F.: SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. In: Proc. EMNLP 2023, pp. 9004–9017. ACL (2023). https://doi.org/10.18653/v1/2023.emnlp-main.557

[6] Min, S., Krishna, K., Lyu, X., et al.: FActScore: fine-grained atomic evaluation of factual precision in long-form text generation. In: Proc. EMNLP 2023, pp. 12076–12100. ACL (2023). https://doi.org/10.18653/v1/2023.emnlp-main.741

[7] Dhuliawala, S., Komeili, M., Xu, J., et al.: Chain-of-Verification reduces hallucination in large language models. arXiv:2309.11495 (2023). https://doi.org/10.48550/arXiv.2309.11495

[8] Gao, L., Dai, Z., Pasupat, P., et al.: RARR: researching and revising what language models say, using language models. In: Proc. ACL 2023, pp. 16477–16508. ACL (2023). https://doi.org/10.18653/v1/2023.acl-long.910

[9] Kadavath, S., Conerly, T., Askell, A., et al.: Language models (mostly) know what they know. arXiv:2207.05221 (2022). https://doi.org/10.48550/arXiv.2207.05221

[10] Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and verification. In: Proc. NAACL-HLT 2018, pp. 809–819. ACL (2018). https://doi.org/10.18653/v1/N18-1074

[11] Wei, J., Wang, X., Schuurmans, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Adv. Neural Inf. Process. Syst., vol. 35, pp. 24824–24837 (2022). https://doi.org/10.48550/arXiv.2201.11903

[12] Lin, S., Hilton, J., Evans, O.: TruthfulQA: measuring how models mimic human falsehoods. In: Proc. ACL 2022, pp. 3214–3252. ACL (2022). https://doi.org/10.18653/v1/2022.acl-long.229

[13] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Adv. Neural Inf. Process. Syst., vol. 30 (2017). https://doi.org/10.48550/arXiv.1706.03762

[14] Rawte, V., Chakraborty, S., Pathak, A., et al.: The troubling emergence of hallucination in large language models. In: Proc. EMNLP 2023, pp. 2541–2573. ACL (2023). https://doi.org/10.18653/v1/2023.emnlp-main.155

[15] Tonmoy, S.M.T.I., Zaman, S.M.M., Jain, V., et al.: A comprehensive survey of hallucination mitigation techniques in large language models. arXiv:2401.01313 (2024). https://doi.org/10.48550/arXiv.2401.01313

[16] Shinn, N., Cassano, F., Gopinath, A., et al.: Reflexion: language agents with verbal reinforcement learning. In: Adv. Neural Inf. Process. Syst., vol. 36 (2023). https://doi.org/10.48550/arXiv.2303.11366

[17] Gou, Z., Shao, Z., Gong, Y., et al.: CRITIC: large language models can self-correct with tool-interactive critiquing. In: Proc. ICLR 2024 (2024). https://doi.org/10.48550/arXiv.2305.11738

[18] National Institute of Standards and Technology: Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1 (2023). https://doi.org/10.6028/NIST.AI.100-1

[19] European Parliament and Council: Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off. J. Eur. Union (2024). http://data.europa.eu/eli/reg/2024/1689/oj

[20] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proc. ICML 2017, pp. 1321–1330. PMLR (2017). https://doi.org/10.48550/arXiv.1706.04599

[21] Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: specialized heads do the heavy lifting, the rest can be pruned. In: Proc. ACL 2019, pp. 5797–5808. ACL (2019). https://doi.org/10.18653/v1/P19-1580
[22] Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proc. ACL 2020, pp. 4190–4197. ACL (2020). https://doi.org/10.18653/v1/2020.acl-main.385
[23] Es, S., James, J., Espinosa-Anke, L., Schockaert, S.: RAGAS: automated evaluation of retrieval augmented generation. arXiv:2309.15217 (2023). https://doi.org/10.48550/arXiv.2309.15217
[24] Xiong, M., Hu, Z., Lu, X., et al.: Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In: Proc. ICLR 2024 (2024). https://doi.org/10.48550/arXiv.2306.13063
[25] Chern, I.C., Chern, S., Chen, S., et al.: FacTool: factuality detection in generative AI — a tool augmented framework for multi-task and multi-domain scenarios. arXiv:2307.13528 (2023). https://doi.org/10.48550/arXiv.2307.13528
[26] Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. In: Proc. ICML 2024. PMLR (2024). https://doi.org/10.48550/arXiv.2305.14325
[27] Hendrycks, D., Burns, C., Basart, S., et al.: Measuring massive multitask language understanding. In: Proc. ICLR 2021 (2021). https://doi.org/10.48550/arXiv.2009.03300
[28] Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Commun. Methods Meas. 1(1), 77–89 (2007). https://doi.org/10.1080/19312450709336664
[29] Chen, W., Su, Y., Yan, X., Wang, W.Y.: KGPT: knowledge-grounded pre-training for data-to-text generation. In: Proc. EMNLP 2020, pp. 8635–8648. ACL (2020). https://doi.org/10.18653/v1/2020.emnlp-main.697
[30] Fomicheva, M., Sun, S., Yankovskaya, L., et al.: Unsupervised quality estimation for neural machine translation. Trans. Assoc. Comput. Linguist. 8, 539–555 (2020). https://doi.org/10.1162/tacl_a_00330